DOCUMENT RESUME

ED 137 372 TM 006 176

| | |
|---|---|
| AUTHOR | Osborn, William C. |
| TITLE | Essential Dimensions of Performance Tests. Professional Paper 1-77. |
| INSTITUTION | Human Resources Research Organization, Alexandria, Va. |
| REPORT NO | HumRRO-PP-1-77 |
| PUB DATE | Apr 77 |
| NOTE | 9p.; Paper presented at the International Congress of Applied Psychology (18th, Montreal, Canada, August 1974) |
| EDRS PRICE | MF-$0.83 HC-$1.67 Plus Postage. |
| DESCRIPTORS | *Performance Tests; Scoring; Testing; *Testing Problems; *Test Reliability; *Test Validity; Training |

ABSTRACT

Four essential dimensions of a performance test are detailed: directness of test method, type of criterion, standardization of conditions, and objectivity of scoring. For simplicity these factors are described as if each were dichotomous, when in actuality each is a continuum; a test method may be more or less direct, conditions more or less standardized. Moreover, as shown here, the dimensions are depicted as independent, when in practice they are not--for instance, indirect methods of testing are often used to attain objective scoring; and process criteria to achieve standardized conditions. Nevertheless, this simple framework provides a useful analytic tool for developers and users of performance tests. It can guide the development of a test, or be used after the fact to identify weaknesses in existing tests. More generally it defines problem areas confronting the performance testing practitioner--problem areas which must be addressed by research and creative development work if performance tests are to be used reliably and validly. (Author/RC)

**********
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
**********

Professional
Paper
1-77

HumRRO-PP-1-77

HumRRO

ED137372

# Essential Dimensions of Performance Tests

William C. Osborn

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Presented at the
18th International Congress of
Applied Psychology
Montreal, Canada August 1974

HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street • Alexandria, Virginia 22314

TM006 116

April 1977

## PREFATORY NOTE

This paper was prepared for, and presented at, the 18th International Congress of Applied Psychology, held in Montreal, Canada, in August 1974. The author, Mr. William C. Osborn, is a Senior Staff Scientist in the HumRRO office at Fort Knox, Kentucky.

# ESSENTIAL DIMENSIONS OF PERFORMANCE TESTS

William C. Osborn

My remarks today are based on a conceptual analysis of factors which constrain the development of valid and reliable performance tests. As one who for several years has been involved in the development of performance tests, I am particularly attuned to the practical problems encountered in trying to provide what might be termed efficient tests—that is, tests which are valid and reliable, but also which are usable in the very real sense of evaluating the proficiency of large numbers of people at minimum cost in time and resources. It is this tradeoff—test quality versus administrative economy—that lies at the heart of the performance testing problem.

Although performance tests have other purposes, they are used chiefly in evaluating training outcomes. Having received training on a job-task (or tasks), a trainee is normally required to demonstrate proficiency on the task before he is advanced to the next stage of training, or ultimately, out of training and onto the job. The development and use of such tests would seem to be straightforward: the job relevant conditions for task performance are created and an acceptable criterion of performance defined. Then the trainee is asked to perform, and his performance is evaluated against the established criterion. Unfortunately, the nature of certain types of job-tasks, together with time and cost constraints, often create problems for the test developer. In circumventing these problems he frequently resorts to simplistic test procedures of questionable reliability or validity. More grave, however, is the fact that such compromises so frequently occur—apparently either because of inadequate regard for the price one pays in diminishing reliability and validity, or because of a lack of awareness of alternate approaches.

My objective today is to set forth in a simple conceptual framework, what I see to be the essential dimensions of a performance test—essential in the sense that they comprise the key practical factors in achieving test reliability and validity. Within this framework I will identify the more common shortcomings of performance tests and then suggest, where I can, possible directions for improvement.

One final caveat before going on: the descriptive model that I will discuss is limited to test development for individual tasks and does not touch on other aspects of reliability and validity—such as sampling of the job task domain or replications of test performance—which pertain to testing on an aggregate of tasks or an entire job.

## TEST METHOD

The first critical dimension of a performance test to be considered pertains to the directness or relevance of what I will call the method of testing. A test method is relevant or direct if it evokes a performance that is the same as that specified in the actual job-task. The scope and fidelity of actual job or life conditions presented and the realism of the response medium used, thus determine the directness of test method. In a training or other performance assessment setting, limited resources often prevent a direct task enactment approach to testing. Indirect methods are often used which involve simulation of task conditions or which require only partial task performance. These commonly result in testing on only part of the task—usually the more testable part. Paper-and-pencil

knowledge tests on tasks with both knowledge and skill requirements represent the most flagrant example of indirect test method. Tests of job knowledge are relatively economical and have exceptional psychometric properties. Yet we would not for a moment consider licensing a man to fly a plane or drive a car, merely on the basis of a knowledge test. The reason for this is obvious. But why then, in other job or job task areas do we tend to accept job knowledge as a valid measure of performance capability? As indicated, the chief reason is cost. A performance test seeks to present the real work environment with all its cues, then elicit the actual job behavior as directly as possible. Such a representation of the real world is expensive. Training and personnel managers tend to think performance tests require too much in the way of equipment, personnel and time to justify their use. But to insist that a test of job knowledge is the only alternative, I believe reflects a false dilemma.

For a given job task several alternate test methods are potentially available. These will lie between an expensive but fully relevant performance test, on the one hand, and a relatively inexpensive but marginally valid knowledge test, on the other. Elsewhere, I have described an approach to devising alternate test methods; an approach based on the concepts of simulation and task-element sampling. Tests resulting from the approach I have collectively termed Synthetic Performance Tests. The intention is to connote a process of synthesis by which the substructure of a job task is used as the basis for selectively constructing alternate forms of a test, each representing (at least theoretically) a more or less optimal blend of validity and feasibility. In some cases this may be achieved through simulation; that is, by substituting for stimuli in either the task display or the surround, or by requiring a substitute response. In other cases, efficient tests may be created by testing on a subset of task elements, regardless of whether simulation is used or not. Thus, synthetically generated alternatives to fully relevant performance tests may vary in two major dimensions, fidelity and scope.

For example, consider an electronic troubleshooting task. Knowing the correct test sequence for isolating a faulty equipment component is only part of the task. Among other task elements the troubleshooter must also be able to place the test-set in operation, establish a good connection at the test points, and correctly interpret the test readouts. Can this type of job task be adequately—that is, validly—tested with the traditional, verbally formatted test of job knowledge? I would say, no. In fact, experience may reveal that, on the job, the most frequent cause of faulty troubleshooting is the inability of the troubleshooter to establish good connections at the test points—an essentially physical or manipulative element in the task performance. So, assuming the test developer cannot afford the luxury of a direct, hand-on method of testing, the important thing is that he does not immediately revert to the typical knowledge test. He should use his inventiveness in devising alternate test methods that will call for the demonstration of behavior that is as similar as possible to that actually required in task performance. Pictorial, graphic, or even low cost three dimensional simulators should be considered. He may then assess the relevance of these synthetic options by checking the breadth and criticality of task elements that are tapped by a particular method.

Only in this way, it seems to me, can test developers arrive at economical methods of proficiency testing while maintaining an acceptable level of content validity.

## TEST CRITERION

Now let me turn to a second dimension of performance tests, that of test criterion. All tasks have both a product (outcome) and process (steps in task performance). Product measurement however, is of overriding importance in certifying a person's achievement on a job task, and failure to include it as the principal criterion may

severely limit test validity. Although it may safely be said that every task has a purpose, the fact of the matter is that in practice a great many performance tests are used which employ process measurement only in evaluating a person's job readiness.

Before looking more closely at why process measures are so widely substituted for measures of task product we must consider three types of tasks. First there are tasks in which the product and the process are one and the same—that is, the product is a process. These tasks are few, and normally are found among those which serve an aesthetic purpose such as springboard diving, dancing, playing a musical composition. Here we see that the outcome or product of the task is no more or less than the correct execution of steps in task performance—that is, the process. A second type of task is that in which the product necessarily follows from the process. Fixed procedure tasks typically fall in this category. Troubleshooting an electrical circuit, balancing a checkbook, changing a tire are examples. In tasks of this type the procedural steps are known, observable and comprise the necessary and sufficient conditions for task outcome; so if the process is correctly executed, task product necessarily follows.

For these first two types of tasks it is not particularly important whether process or product measurement is used. But for a third type it is. This is the type in which the product is less than fully predictable from the process—a circumstance which occurs either because we are unable to fully specify the necessary and sufficient steps in task performance, or because we cannot or do not accurately measure them. In spite of the obvious importance of product measurement for tasks in this latter category, in practice performance tests often do not focus on product. And the reasons generally stem from practical considerations in which the measurement of task product is viewed as too costly, too dangerous, or for other reasons simply too impractical. For example, in a first aid task involving controlling the bleeding from an external wound, the test developer would probably be limited to requiring demonstration of task process; observation of the actual task product—restriction of blood flow—would probably not be possible, for obvious reasons. Other situations are less understandable. If any of you are involved in the field of instructor training, you may have observed that a student instructor is evaluated on the basis of such process factors as: "had a well organized lesson plan," "used visual-aids effectively," "had good eye contact," "had good voice projection," "covered all points in the lesson plan," etc. Although clearly the product of instruction is student learning, I believe it is seldom, if ever, used as the criterion for qualifying an instructor—probably because it would involve a more time consuming method of evaluation.

I'm sure we could all testify to other instances in which product measurement is not used. Some of these are justified by cost or safety considerations, but others are not. It seems to me that test developers often fail to see the importance of measuring task outcome; or perhaps they merely slight its importance when faced with practical limitations in its measurement. The overriding question that a test designer should ask himself in this situation is, "If I use only a process measure to test a person's achievement on a task, how certain can I be from this process score that the person would also be able to effect the product or outcome of the task?" Where the degree of certainty is substantially less than that to be expected from normal measurement error, the test designer should pause and reconsider ways in which time and resource limitations can be compromised in achieving at least an approximation to product measurement.

## TEST CONDITIONS

Now, let's look at a third dimension of performance tests—that of standardization of conditions under which a test is administered. This is an important step in achieving test reliability. Indeed, the very essence of any proficiency measure which professes to

be a test, is that of standardized conditions. This requirement is familiar to test developers and is therefore less often violated. An effort is normally made to maintain test instructions, materials, tools, and other environmental factors as nearly constant as possible from one test administration to the next. However, I would like to call to your attention one particular class of tasks which is particularly troublesome in this regard: tasks involving interpersonal behavior. Here, another person or group of people represent an important part of the environment to be controlled—that is standardized—from one test administration to the next. Examples are seen in such areas as counseling, salesmanship, personnel management, or in something like hand-to-hand combat. Tasks in these areas all entail other people as part of the task relevant conditions; and obviously people are difficult to standardize. If you were interested in assessing a policeman's ability to properly subdue an unarmed but hostile suspect, what would your performance test be like? And how would you insure that test conditions were standardized over all policemen to be tested? The same question might be asked in relation to assessing a would-be supervisor's ability to persuade a worker to perform some difficult or unpleasant task.

Unfortunately, I know of no easy solution to this problem. Probably, the direction that test designers should take is toward greater use of the well trained, "standardized other" in controlled role-playing situations. In any case, the product in these kinds of tasks is some defined, observable change in that task-relevant "other." And, here, greater effort should be made to avoid settling too quickly for some probably irrevalent measure of task process.

## TEST SCORING

The fourth and final dimension essential to performance tests is that of test scoring. Scoring-protocols impact primarily on reliability, but if grossly mishandled in test design, as I will point out in a moment, they may also jeopardize test validity. Scoring procedures involve translating an observed test outcome into an objective pass-fail score. Such procedures should be structured so that only the more reliable perceptual skills are used; that is, the scoring activity should be reduced to one of matching or comparing the test response with some model of correct response. Unfortunately responses in many test situations seemingly cannot be judged in this "either or" fashion, but require a "more-or-less" type of judgment. When this occurs the test developer should not, as is sometimes done, escape by using a test method that yields a more measurable outcome, because test validity may suffer. Rather, he should remain with the task-relevant response and strive to break it down into elements so that comparative judgments can be made more easily by a scorer. A familiar illustration of what I mean is seen in typical programs of knowledge testing. The pervasive multiple-choice test yields responses which can be scored with maximum reliability. Obviously, scorers have little difficulty in matching a selected response alternative with that which is keyed as correct by the test developer. The scoring of essay tests, on the other hand, has traditionally presented reliability problems. Yet in spite of the scoring problems inherent in essay testing, the competent test developer would not resort to multiple-choice testing on knowledge tasks demanding recall or generation of material merely to achieve greater scorer reliability. Normally he would provide a model response in the form of an exhaustive list of the critical elements of an acceptable essay response, the presence of which can be judged with relative objectivity by a qualified and earnest scorer.

This same thinking applies to the development of scoring protocols for performance tests if these tests are to produce reliable results. The subjectivity with which many task performances are customarily scored could be substantially reduced, it seems to me, through wider use of what may be termed scoring templates. Where the model response

on a test of marksmanship is defined as a hole in the bullseye, it is relatively easy for the scorer to judge the acceptability of the response made by the rifleman. This is because the concentric circles normally marked on a target act as a kind of simple template which enhances the ease and objectivity of scorer judgments as to the nearness of a hit to the center of the target. Templates could be applied equally well in scoring other tests. For example, tasks mentioned earlier in which the outcome is a process are often troublesome to assess reliably. It would appear that performances such as springboard diving or gymnastic exercises could be more objectively scored if the outcomes were filmed and figural templates overlayed on key frames to assess the accuracy of the performer at those critical points in the response. Similarly, in evaluating the performance of a music student, recordings of selected renditions could be analyzed at the scorer's leisure perhaps with the aid of auditory templates such as a metronome to measure beat or comparative tones to assess tonal quality. For these particular tasks—or for that matter, any task in which the product is transient—the added cost in recording the product for scoring later would probably be offset by savings in scoring costs; that is, the more objective approach to scoring would very likely preclude the usual requirements for a panel of expert evaluators. But more importantly the scorer would not be constrained by real time, and could function at a place and time and rate of his choosing, using prepared templates to further the objectivity of his judgments.

Thus we have what I consider to be the four essential dimensions of a performance test: directness of test method, type of criterion, standardization of conditions, and objectivity of scoring. For simplicity these factors have been described as if each were dichotomous, when in actuality each is a continuum; a test method may be more or less direct, conditions more or less standardized. Moreover, as shown here, the dimensions are depicted as independent, when in practice they are not—for instance, indirect methods of testing are often used to attain objective scoring; and process criteria to achieve standardized conditions.

Nevertheless, this simple framework provides a useful analytic tool for developers and users of performance tests. It can guide the development of a test, or be used after the fact to identify weaknesses in existing tests. More generally it identifies problem areas confronting the performance testing practitioner—problem areas which must be addressed by research and creative development work if performance tests are to be used validly and reliably.

## Essential Dimensions of Performance Tests